Seminario di Sicurezza - Filtri Antispam

Daniele Venzano

27 dicembre 2003

Sommario

Il problema dello SPAM sta diventando di proporzioni sempre più grandi. Alcuni Service Provider riportano che l'80% del traffico SMTP sulle loro reti è provocato da messaggi indesiderati di vario tipo.
In questo seminario si dà un'idea di quali sono le tecnologie correnti per il filtraggio automatizzato dei messaggi SPAM. Verranno anche discusse le differenze di una soluzione lato server rispetto ad una lato client.

1 Metodi di riconoscimento dello SPAM

I messaggi di SPAM sono generati automaticamente ed in grande quantità, motivo per cui hanno una serie di caratteristiche particolari che facilitano il compito di un filtro automatico. In particolare i messaggi tendono ad usare un sottoinsieme ristretto di parole e spesso usano molto HTML (con colori brillanti, e font grandi).

1.1 Filtri basati su hashing dei messaggi

Questo tipo di filtri riconoscono tutti i messaggi di SPAM uguali successivi al primo. Esistono due tipi di implementazioni differenti, ma in generale hanno prestazioni intorno al 50-70% di messaggi riconosciuti, e raramente hanno falsi positivi. Sono vulnerabili ad attacchi di *poisoning*.

1.1.1 Sistemi collaborativi

Un tipo di implementazione (Razor, Pyzor) è distribuita e collaborativa ed è adatta sia a server che a client. Quando un messaggio viene riconosciuto come SPAM usando uno dei metodi elencati sotto, ne viene generato un codice hash che poi viene inviato ad un server.
Per controllare un messaggio, semplicemente si controlla se il suo codice hash è elencato sul server.

Questo tipo di riconoscimento è facilmente aggirabile aggiungendo ai messaggi di SPAM delle combinazioni casuali di lettere e numeri, diversi per ogni messaggio. I filtri si tengono alla pari calcolando l'hash solo su pezzi particolari del messaggio, ma è una continua rincorsa tra chi scrive i messaggi di spam e chi scrive i filtri.

1.1.2 Sistemi chiusi

Questo tipo di sistemi sono adatti a server che lavorano su ampi volumi di posta e quindi non si possono permettere di aspettare il tempo necessario a verificare il codice hash su un server remoto per ogni mail di passaggio. Quindi funzionano mantenendo una cache locale di codici hash.

Soffrono degli stessi problemi dei sistemi collaborativi e in più la loro efficienza è ulteriormente diminuita dall'insieme ristretto di messaggi su cui lavorano (e quindi delle cache di codici hash).

1.2 Filtraggio Bayesiano (o Statistico)

I filtri bayesiani sono considerati, attualmente, i migliori. Quando un messaggio viene dato in pasto ad un filtro bayesiano, ad ogni parola viene assegnata una probabilità calcolata in base ai messaggi precedenti.

I filtri bayesiani hanno bisogno di un periodo di training in cui gli vengono forniti sia messaggi SPAM, che messaggi normali. In questo periodo il filtro si costruisce una tabella di probabilità per tutte le parole che incontra. Per questa fase di training c'è bisogno, di solito, di qualche centinaio di messaggi di entrambi i tipi. Una volta che il filtro ha imparato a riconoscere i messaggi e viene messo 'in produzione', continua ad aggiornare le sue tabelle, in modo da adattarsi automaticamente a nuovi tipi di messaggi.

Filtri di questo tipo riescono a riconoscere anche più del 95% dei messaggi SPAM, con una percentuale di falsi positivi decisamente bassa, inferiore al 1%.

Normalmente i filtri bayesiani non vengono utilizzati da soli, ma sono sempre associati ad altri sistemi di riconoscimento per diminuire al massimo la possibilità di falsi positivi.

Tra i metodi utilizzati per cercare di ridurre la loro efficacia, c'è quello di scrivere le parole in modo da renderle irriconoscibili ad un tool automatico (ad esempio 's–e–x' o 'Order H–E–R–E'). Questo ha l'effetto opposto, in quanto per un filtro bayesiano, dopo le prime due o tre mail, le parole che contengono molti '-', diventano un sicuro indice di spam. Nei messaggi normali quelle parole non compaiono mai.
Metodi di quel tipo, però, possono dare problemi ai filtri basati su punteggi, anche se è facile per chi scrive i filtri rimanere al passo.

1.3 Filtraggio basato su punteggi (o Euristico)

Questo tipo di filtri compie una serie di controlli su ogni messaggio, assegnando un punteggio ad ogni regola che viene verificata. Alla fine si fa la somma di tutti i punteggi accumulati e se questa supera una certa soglia, il messaggio viene marcato come SPAM.

I controlli sono sia di integrità della mail, come corrispondenza tra gli header (From: e il primo Received:, ad esempio) o correttezza delle informazioni MIME, che di ricorrenza di certe parole (#FF0000, il codice HTML per il rosso è considerata una delle parole più significative da moli filtri). Ma nulla vieta che siano considerati come punteggi anche risultati di filtri bayesiani o basati su codici hash.

1.3.1 Spamassassin

Spamassassin è considerato, attualmente, il migliore sistema antispam, ed è un filtro a punteggi che comprende anche, come controlli aggiuntivi, un filtro bayesiano, diversi sistemi di blacklist e due filtri collaborativi basati su hashing (Pyzor e Razor).

Spamassassin è pensato per essere utilizzato indifferentemente su un sistema server o su un client, infatti quando un messaggio SPAM viene riconosciuto, non viene scartato, ma soltanto marcato in modo speciale. É l'utente finale sul suo programma di posta che decide quanti punti un messaggio deve avere per essere scartato.

Un messaggio trattato da Spamassassin e riconosciuto come SPAM viene modificato aggiungendo degli header che mostrano quali test sono risultati positivi e quale è il punteggio complessivo.

```
X-Spam-Flag: YES
X-Spam-Checker-Version: SpamAssassin 2.60 (1.212-2003-09-23-exp)
X-Spam-Status: Yes, hits=5.5 required=4.0 tests=BAYES_90,HTML_MESSAGE,
        NORMAL_HTTP_TO_IP,SUB_HELLO,TO_ADDRESS_EQ_REAL autolearn=no
        version=2.60
X-Spam-Level: *****
```

Il messaggio originale viene allegato ad una nuova mail che contiene l'analisi del messaggio SPAM:

```
Content analysis details:   (5.5 points, 4.0 required)

 pts rule name              description
---- ---------------------- --------------------------------------------------
 2.5 SUB_HELLO              Subject starts with "Hello"
 0.8 TO_ADDRESS_EQ_REAL     To: repeats address as real name
 2.1 BAYES_90               BODY: Bayesian spam probability is 90 to 99%
                            [score: 0.9274]
 0.1 HTML_MESSAGE           BODY: HTML included in message
 0.1 NORMAL_HTTP_TO_IP      URI: Uses a dotted-decimal IP address in URL
```

Il messaggio che segue è interessante perchè comprende alcune delle tecniche utilizzate dagli autori di SPAM per circuire i filtri. Queste tecniche gli permettono di avere un punteggio basso, ma non abbastanza per passare il filtro.
Bisogna notare, comunque, che il messaggio è molto breve (0,7KB di testo utile). Infatti più il testo è lungo, più è probabile che il filtro riconosca il messaggio per quello che è. D'altra parte un messaggio troppo breve non riesce a trasmettere un messaggio commerciale convinciente.

```
Return-Path: <enwo@mail.ru>
Delivered-To: venza@localhost
Received: from pop.libero.it [193.70.192.70]
        by localhost with POP3 (fetchmail-5.9.11)
        for venza@localhost (single-drop); Wed, 22 Nov 2003 22:14:41 +0200 (MEST)
Received: from smtp5.libero.it (193.70.192.55) by ims4c.libero.it (7.0.019)
        id 3F3B8EBB01486453 for venza@libero.it; Wed, 22 Nov 2003 22:10:15 +0200
Received: from pD9E00E25.dip.t-dialin.net (217.224.14.37) by smtp5.libero.it (7.0.020-DD01)
        id 3F60363804B5FC6C for venza@libero.it; Wed, 22 Nov 2003 22:10:15 +0200
Received: from FHCBADIAC ([192.168.185.226])
        by pD9E00E25.dip.t-dialin.net (8.11.6/8.11.6) with SMTP id h4331CB07160
        for <venza@libero.it>; Wed, 22 Nov 2003 20:12:23 +0000
Message-ID: <270f01c398d8$64dc142e$496ecbcf@FHCBADIAC>
From: "Anna hryic" <enwo@mail.ru>
To: "venza@libero.it" <venza@libero.it>
Subject: Hello, it's again me.
Date: Wed, 22 Nov 2003 20:12:00 +0000
X-Priority: 3
```

```
X-MSMail-Priority: Normal
X-Mailer: Microsoft Outlook Express 6.00.2800.1158
X-MimeOLE: Produced By Microsoft MimeOLE V6.00.2800.1165
X-UID: 979
```

Gli header sono verosimili e non contengono incogruenze immediatamente visibili, anche se probabilmente sono inventati. Chi manda SPAM, infatti, ne manda qualche milione ed usa dei programmi (o script) fatti apposta per inviare grossi volumi di posta. Non usano di certo Outlook Express.
Il *subject* che inizia con la parola Hello ha contribuito per quasi la metà del punteggio totale.

```
Content-Type: text/plain;

n dneopuu pifrb iati a panrfi npei eymayiu rp xerdee
```

Questo è il messaggio in testo semplice, è una stringa di caratteri casuali, probabilmente diversa per ogni messaggio. É pensata per raggirare i filtri basati su hashing. Molti client di posta 'intelligenti' mostrano prima il testo in HTML e solo se questo non è presente, mostrano quello in solo testo.

```
Content-Type: text/html;

Good day!
I's again me, i wrote you few days ago, do you remember me?
I live in a small city and i do not meet very many people, i found
great site, look at it
http://218.15.192.225/index.php?a=3D000344&b=3D001&c=3D20
also "http://218.15.192.225/ugly.php?campaign=3D000344
there you can ask me not send more emails to you.

Anna ssltame
```

Il testo non contiene parole particolarmente evidenti (come sex, teens, unsubscribe, ecc.), ma il filtro bayesiano (basandosi anche sugli header) gli ha assegnato lo stesso una probabilità del 92% di essere un messaggio SPAM. Il cognome ssltame è, di nuovo, casuale, e probabilmente diverso per ogni mail.
Il codice 3D000344, codificato nell'URL, serve, probabilmente, ad identificare l'indirizzo del destinatario, in modo da confermare al mittente che è valido non appena il link viene seguito.
L'indirizzo IP appartiene ad un provider cinese (China Telecom), ma non è già più raggiungibile. Se chi si fa pubblicità con lo SPAM è costretto a cambiare indirizzi troppo velocemente, perde anche quei pochi possibili clienti, e il sistema diventa antieconomico.

1.4 Filtraggio Challenge-Response

Questo tipo di filtri sono utilizzati lato utente e sono adatti a chi non riceve posta da nuovi indirizzi frequentemente.

Ogni volta che arriva una mail da un indirizzo sconosciuto, viene risposto in maniera automatica un nuovo messaggio che chiede di confermare l'invio.
Dato che nella maggior parte dei casi i mittenti delle mail SPAM sono falsificati, questo sistema ha un'ottima resa. Il problema è che si sposta l'onere del filtraggio a chi spedisce i messaggi regolari.

1.5 Mail Server Blacklists

Le blacklist sono elenchi di indirizzi IP conosciuti come fonte di SPAM e sono utilizzate lato server per rifiutare le connessioni dai server di posta elencati. Anche se l'idea può sembrare buona, una delle blacklist più usate (MAPS RBL) ha delle percentuali di resa non proprio felici: SPAM riconosciuto 24%, falsi positivi 34%.

Questo è causato dal fatto che è molto facile essere inseriti in una blacklist, ed è, al contrario, molto difficile uscirne. In più molte blacklist sono usate come un sistema punitivo verso i provider più negligenti, inserendo intere classi di IP, anche se lo SPAM arriva da un solo indirizzo. Questo, sì, constringe i provider ad agire contro chi manda messaggi SPAM, ma blocca, alle volte anche per qualche settimana, tutta la posta leggitima proveniente dal provider inserito nella lista, provocando danni ingenti a persone del tutto estranee.

2 Eliminazione dello SPAM

Una volta che un messaggio è stato riconosciuto come SPAM, come è meglio comportarsi ?
La risposta non è unica e dipende da dove il messaggio si trova (server o client) e da quanto è costosa la perdita, per quanto rara, di un messaggio marcato erroneamente come SPAM.

Su un server, eliminare i messaggi indesiderati ha dei grossi vantaggi, riduce lo spreco di banda ed elimina virus e trojan che potrebbero essere scaricati da utenti inconsapevoli che abboccano ai messaggi pubblicitari. D'altra parte il costo per un messaggio legittimo perso è alto perchè processando un grosso volume di posta la probabilità di scontentare qualcuno perdendogli un messaggio è alta. Questa potebbe essere una buona soluzione per una rete aziendale, dove gli utenti, spesso, non hanno le conoscenze necessarie ad impostare dei filtri lato client.

Invece una soluzione per un ISP o un fornitore di indirizzi email potrebbe essere quella di spezzare il processo di filtraggio: sul server si fa l'identificazione e sul client l'utente decide che fare dei messaggi riconosciuti. Viene a mancare il risparmio di banda e gli utenti devono sapere come fare a filtrare i messaggi indesiderati. In compenso ogni utente può scegliere a quale livello iniziare a filtrare automaticamente. Si può anche filtrare oltre un certo livello direttamente sul server, in modo da eliminare i casi più evidenti.
In questo modo il filtraggio diventa un servizio in più fornito all'utenza, che può decidere se usufruirne oppure no.

3 Bibliografia

1. http://www.paulgraham.com/spam.html
2. http://www.spamassassin.org